DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays

Amrit Singh[1], Casey P. Shannon[1], Benoît Gautier[2], Florian Rohart[3], Michaël Vacher[4], Scott J. Tebbutt[1] and Kim-Anh Lê Cao[5]

[1]Prevention of Organ Failure (PROOF) Centre of Excellence, University of British Columbia, Vancouver, BC, Canada.
[2]The University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, QLD 4102, Australia
[3]Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia
[4]Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia
[5]Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia

Section S1: Simulated datasets

Description of simulation analysis, from generating synthetic multi-omics data to applying various integrative classification approaches.

Simulated datasets

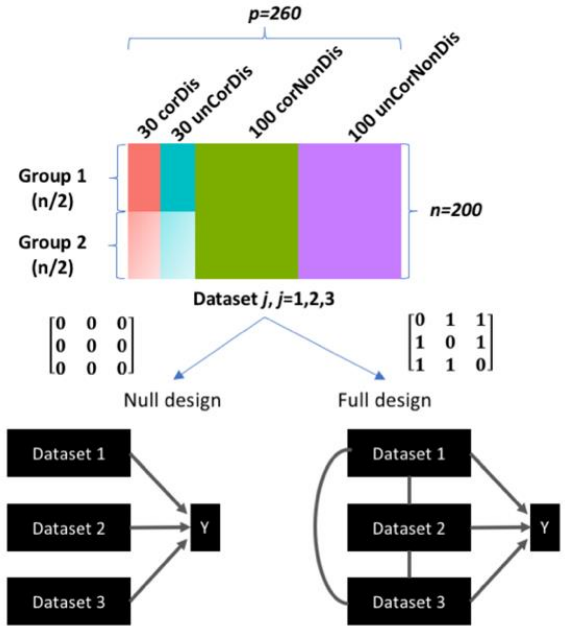

Figure. Simulated multi-omics data. Each simulated dataset consisting of four types of variables: 30 correlated-discriminatory (corDis) variables, 30 uncorrelated-discriminatory (unCorDis) variables, 100 correlated-nondiscriminatory (corNonDis) variables and 100 uncorrelated-nondiscriminatory (unCorNonDis) variables.

Three datasets were simulated each with 200 observations (n) and 260 variables (p). The 200 observations were split equally into two groups (G1 and G2), whereas the 260 variables were generated by varying the covariance ($cov(X_i, X_j) = [0,5,10,15],\ where\ i \neq j$, between datasets and fold-change ($\Delta = \mu_{G2} - \mu_{G1} = [0,1,2]$) between G1 and G2: 30 correlated-discriminatory (corDis) variables, 30 uncorrelated-discriminatory (unCorDis) variables, 100 correlated-nondiscriminatory (corNonDis) variables, and 100 uncorrelated-nondiscriminatory (unCorNonDis) variables were simulated (Figure left). The resulting dataset was of the form:

$X_j = [\, X_j^{corDis} \mid X_j^{unCorDis} \mid X_j^{corNonDis} \mid X_j^{unCorNonDis} \,] + E_j$, where j = 1, 2, 3

Correlated and discriminatory variables, X_j^{corDis} (200 samples x 30 variables per dataset j)

The matrix containing correlated and discriminatory variables, X_j^{corDis} was generated using the following model:

$X_j^{corDis} = \boldsymbol{u}_j^{corDis} \boldsymbol{w}_j^t$, where $\|\boldsymbol{w}\| = 1, j = 1,2,3$

where the loadings, $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ were vectors of length 30, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] $\cup$ [0.2, 0.3]. For G1 (G2), the outer components $\boldsymbol{u}_1^{corDis}$, $\boldsymbol{u}_2^{corDis}$, $\boldsymbol{u}_3^{corDis}$ were vectors of length 100 drawn from a multivariate normal distribution with a mean value of $-\Delta/2$ ($\Delta/2$), where the grid values of 0, 1, 2 were used for Δ. The covariance between pairs of components, cov($\boldsymbol{u}_i^{corDis}$, $\boldsymbol{u}_j^{corDis}$) was set to 1 for all *i, j=1,2,3.*

Uncorrelated and discriminatory variables, $X_j^{unCorDis}$ (200 samples x *30 variables per dataset j)*
The matrix containing uncorrelated and discriminatory variables, $X_j^{unCorDis}$ was generated using the following model:
$X_j^{unCorDis} = \boldsymbol{u}_j^{unCorDis}\boldsymbol{w}_j^t$, where $\|\boldsymbol{w}\| = 1, j = 1,2,3$
where the loadings, $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ were vectors of length 30, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] ∪ [0.2, 0.3]. For G1 (G2), the outer components $\boldsymbol{u}_1^{unCorDis}$, $\boldsymbol{u}_2^{unCorDis}$, $\boldsymbol{u}_3^{unCorDis}$ were vectors of length 100 drawn from a multivariate normal distribution with a mean value of $-\Delta/2$ ($\Delta/2$), where the grid values of 0, 1, 2 were used for Δ. The covariance between pairs of components, cov($\boldsymbol{u}_i^{unCorDis}$, $\boldsymbol{u}_j^{unCorDis}$) was set to 0 when $i \neq j$ and 1 when $i = j$.

Correlated and nondiscriminatory variables, $X_j^{corNonDis}$ (200 samples x *100 variables per dataset j)*
The matrix containing correlated and nondiscriminatory variables, $X_j^{corNonDis}$ was generated using the following model:
$X_j^{corNonDis} = \boldsymbol{u}_j^{corNonDis}\boldsymbol{w}_j^t$, where $\|\boldsymbol{w}\| = 1, j = 1,2,3$
where the loadings, $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ were vectors of length 100, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] ∪ [0.2, 0.3]. The outer components $\boldsymbol{u}_1^{corNonDis}$, $\boldsymbol{u}_2^{corNonDis}$, $\boldsymbol{u}_3^{corNonDis}$ were vectors of length 200 drawn from a multivariate normal distribution with a mean value of 0. The covariance between pairs of components, cov($\boldsymbol{u}_i^{corNonDis}$, $\boldsymbol{u}_j^{corNonDis}$) was set to 1 for all *i, j=1,2,3.*

Uncorrelated and nondiscriminatory variables, $X_j^{unCorNonDis}$ (200 samples x *100 variables per dataset j)*
The matrix containing uncorrelated and discriminatory variables, $X_j^{unCorNonDis}$ was generated using the following model:
$X_j^{unCorNonDis} = \boldsymbol{u}_j^{unCorNonDis}\boldsymbol{w}_j^t$, where $\|\boldsymbol{w}\| = 1, j = 1,2,3$
where the loadings, $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ were vectors of length 100, and the elements were drawn from a uniform distribution in the interval of [-0.3, 0.2] ∪ [0.2, 0.3]. The outer components $\boldsymbol{u}_1^{unCorNonDis}$, $\boldsymbol{u}_2^{unCorNonDis}$, $\boldsymbol{u}_3^{unCorNonDis}$ were vectors of length 200 drawn from a multivariate normal distribution with a mean value of 0. The covariance between pairs of components, cov($\boldsymbol{u}_i^{unCorDis}$, $\boldsymbol{u}_j^{unCorDis}$) was set to 0 when $i \neq j$ and 1 when $i = j$.

The residual matrix, $\boldsymbol{E}_j$ is a 200 x 260 residual matrix where each element is drawn from a normal distribution with zero mean and variance equal to 0.2, 0.5, or 1.

Simulation analysis

Holding covariance constant at 1 (Figure 1 in main manuscript)
Using a fold-change grid of [0, 1, 2] and noise grid of [0.2, 0.5, 1], sets of three datasets were simulated for each fold-change and noise combination. Then a DIABLO model was generated using either the full or null design (DIABLO_full and DIABLO_null). One component was

retained in the DIABLO model, selecting 60 variables from each dataset for a total of 180 variables (across all datasets). In addition, other integrative schemes such as concatenation and ensemble-based classifiers were also tested using the sPLSDA classifier. For the concatenation-based scheme, all datasets were concatenated into one matrix containing 3x260=880 variables and sPLSDA was applied, retaining 1 component and 90 variables. For the ensemble-based scheme, a sPLSDA classifier was applied to each dataset separately retaining one component and 30 variables per dataset. The consensus predictions were determined using a majority vote scheme. A 10-fold cross-validation averaged over 20 simulations was used to evaluate the performance of each method/scheme and the number of each type of variable selected in each model was recorded.

Holding noise constant at 0.5 (Supplementary Figure S2 below)
Using a fold-change grid of [0.5, 1, 2, 4] and a covariance grid of [0, 5, 10, 15], sets of three datasets were simulated for each fold-change and covariance combination. For each combination, a DIABLO model with either the full or null design were generated, and the error rate was evaluated using a 10-fold cross-validation. This procedure was repeated 20 times and an average error rate for determined. For Supplementary Figures S2 A-B, the DIABLO models consisted of 1 component, retaining 60 variables per component per dataset (180 variables in total) whereas for Supplementary Figures S2 C-D, the DIABLO models consisted of 2 components, retaining 30 variables per component per dataset (180 variables in total).

Section S2: Real world datasets.

Details regarding the multi-omics data used for the benchmarking experiments and case studies (breast cancer and asthma).

Benchmarking cancer datasets

All cancer (colon, glioblastoma, kidney and lung) datasets used for the benchmarking analyses were obtained from http://compbio.cs.toronto.edu/SNF/SNF/Software.html (Wang *et al.*, 2014). For the mRNA datasets, all transcripts with the same gene symbol were averaged.

Breast cancer multi-omics study

Datasets accession: The level 3 TCGA data (version 2015_11_01) were retrieved from firebrowse.org hosted by the Broad Institute. The clinical data file (Merge_Clinical) was downloaded from the Primary tab of the BRCA Clinical Archives. The mRNA RSEM normalized dataset (illuminahiseq_rnaseqv2-RSEM_genes_normalized) was downloaded from the Primary tab of the BRCA mRNASeq Archives. The miRNA datasets (illuminahiseq_mirnaseq-miR_gene_expression and illuminaga_mirnaseq-miR_gene_expression) were downloaded from the Primary tab of the BRCA miRSeq Archives. The reverse phase protein array dataset (mda_rppa_core-protein_normalization) was downloaded from the Primary tab of the BRCA RPPA Archives. The beta values for the methylation datasets (humanmethylation27-within_bioassay_data_set_function and humanmethylation450-within_bioassay_data_set_function MD5) were downloaded from the Primary tab of the BRCA Methylation Archives.

Data processing: Clinical data were present for 1,098 subjects for 3,703 variables. 29 unannotated transcripts were removed from the mRNA dataset composed resulting in 20,502 genes x 1212 samples. Two transcripts corresponded to *SLC35E2*, therefore one of the transcripts was re-

labelled *SLC35E2.rep*. The miRNA datasets (1,046 miRNA x 1190 samples) was derived using two different Illumina technologies, the Illumina Genome Analyzer (341 samples) and the Illumina HiSeq (849 samples). The read counts instead of the reads_per_million_miRNA_mapped were used. The proteomics dataset obtained using a reverse phase protein array consisted of 142 proteins for 410 samples. The methylation data was derived from two different platforms, the Illumina Methylation 27 (27,578 CpG probes x 343 subjects) and the Illumina 450K (485,577 CpG probes x 885 subjects). There were 25,978 CpG probes in common between the platforms. The PAM50 labels for 1,182 samples were obtained from the TCGA staff. All datasets were restricted to samples coming from the primary solid tumor (sample type code 01) and to the first vial (vial code A).

Normalization and pre-filtering: The count data for the mRNA dataset, X_{counts} was normalized to log2-counts per million (logCPM), X_{norm}, similar to limma voom (Law *et al.*, 2014):

$$X_{norm} = \log_2\left(\frac{\left(X_{counts} + 0.5\right)^T}{\left(lib.size + 1\right) * 10^6}\right)$$

After library size (lib.size = total number of reads per sample) normalization, genes with counts less than 0 in more than 70% of samples were removed. The PAM50 genes were also removed from the mRNA dataset prior to analyses. Similarly, the miRNA count data was normalized to logCPM and miRNA transcripts with counts less than 0 in more than 70% of the samples were also removed.

Asthma multi-omics study

Datasets accession: Paired blood samples were obtained from 14 asthmatic individuals undergoing allergen inhalation challenge as previously described (Singh *et al.*, 2012). Cell counts were obtained from a hematolyzer (percentage of Neutrophils, Lymphocytes, Monocytes, Eosinophils and Basophils) and DNA methylation analysis (percentage of T regulatory cells, T cells, B cells and Th17 cells). Gene expression profiling was performed using Affymetrix Human Gene 1.0 ST (GSE40240). Metabolite profiling was performed by Metabolon Inc. (Durham, North Carolina, USA). All asthma data have been published as part of previous studies (Singh *et al.*, 2013, 2014).

Normalization: Microarray data was normalized using Robust MultiArray Average (RMA), consisting of background correction, quantile normalization and probe summarization using median polish. Preprocessing of mass spectrometry data including data extraction, peak-identification and data preprocessing for quality control and compound identification was performed by Metabolon Inc. (Durham, North Carolina, USA).

Section S3: Description of methods used for the benchmarking experiments.

Parameters settings used for the various integrative approaches applied to the benchmarking cancer datasets.

Description of methods used for the benchmarking experiments

For the purposes of this study, only component-based methods that integrated multiple datasets and perform variable selection were considered. Since tuning the number of variables to retain in each model would result in biomarker panels with different numbers of variables, for the purposes of this study all variables were retained in each model. The features were instead ranked based on their absolute value of their loadings (importance) and 60 variables were selected from each omic type, resulting in multi-omic biomarker panels with 180 variables (60 mRNAs, 60 miRNAs and 60 CpGs). Equal numbers of variables allowed for a fair comparison in the gene set enrichment analysis.

	Parameter settings
Supervised	
DIABLO_null	ncomp = 2 (# of components) keepX = all variables were retained from each omics dataset $design = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ default parameters were used for the other arguments: scheme="horst", mode="regression", scale = TRUE, init = "svd", tol = 1e-06, max.iter = 100
DIABLO_full	ncomp = 2 (# of components) keepX = all variables were retained from each omics dataset $design = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ default parameters were used for the other arguments: scheme="horst", mode="regression", scale = TRUE, init = "svd", tol = 1e-06, max.iter = 100
Concatenation-sPLSDA	ncomp = 2 (# of components) keepX = all variables were retained from each omics dataset default parameters were used for the other arguments: mode = "regression"

	scale = TRUE, tol = 1e-06, max.iter = 100
Ensemble_sPLSDA	ncomp = 2 (# of components) keepX = all variables were retained from each omics dataset default parameters were used for the other arguments: mode = "regression" scale = TRUE, tol = 1e-06, max.iter = 100
Unsupervised	
sGCCA (Tenenhaus *et al.*, 2014)	ncomp = 2 (# of components) keepX = all variables were retained from each omics dataset $design = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ default parameters were used for the other arguments: scheme = "horst", mode="canonical", scale = TRUE, init = "svd.single", tol = .Machine$double.eps, max.iter=1000,
JIVE*(Lock *et al.*, 2013)	default parameter settings from the jive() from the r.jive R-package were used: 1. scale = TRUE, center = TRUE 2. method = "perm" sPCA parameters: ncomp = 2 (# of components) keepX = rep(ncol(X),ncomp)(all variables were retained from each omics dataset default parameters were used for the other arguments: center = TRUE scale = TRUE, max.iter = 500, tol = 1e-06
MOFA (Argelaguet *et al.*, 2018)	factors=2 (# of components) default parameter settings recommended by MOFA were used: 1. likelihoods=(gaussian gaussian gaussian) 2. Convergence criterion (tolerance=0.01, nostop=0)

	3. Training components (startDrop=1 # initial iteration to start shutting down factors, freqDrop=1 # frequency of checking for shutting down factors, dropR2=0.00 # threshold on fraction of variance explained) 4. hyperparameters for the feature-wise spike-and-slab sparsity prior [learnTheta=(1 1 1) # 1 means that sparsity is active whereas 0 means the sparsity is inactivated; each element of the vector corresponds to a view, initTheta=(1 1 1) # initial value of sparsity levels (1 corresponds to a dense model, 0.5 corresponds to factors); each element of the vector corresponds to a view, startSparsity=250 # initial iteration to activate the spike and slab, we recommend this to be significantly larger than 1] Intercept was set to TRUE (learnIntercept=1)

*since the variable selection functionality has not been added to JIVE R-function, sparse Principal Component Analysis (sPCA) from the mixOmics R-package was applied to the joint variation matrix obtained after applied JIVE to the multi-omics cancer datasets.

Section S4: Gene-set enrichment analyses

Significance of enrichment was determined using a hypergeometric test of the overlap between the selected features (mapped to official HUGO gene symbols or official miRNA symbols) and the various gene sets contained in the collections. The false discovery rate was computed for each collection separately using the Benjamini Hochberg False Discovery Rate (Benjamini and Hochberg, 1995) procedure. The number of gene sets with an FDR less than 5% were determined and used as a metric to compared different multi-omics integrative methods.

In order to carry out the comparison, each feature set was mapped back to official HUGO gene symbols. This was done as follows across the respective data types: mRNA, CpGs and proteins (when present). The following collections were used as gene-sets for the enrichment analysis (Subramanian *et al.*, 2005): C1 - positional gene sets for each human chromosome and cytogenetic band. C2 – curated gene sets (Pathway Interaction DB [PID], Biocarta [BIOCARTA], Kyoto Encyclopedia of Genes and Genomes [KEGG], Reactome [REACTOME], and others), C3 - motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes. C4 – computational gene sets (from the Cancer Gene Neighbourhoods [CGN] and Cancer Modules [CM] – citation available via the MolSigDB (Liberzon *et al.*, 2015). C5 - GO gene sets consist of genes annotated by the same GO terms. C6 – ontologic gene sets (Gene sets represent signatures of cellular pathways which are often dis-regulated in cancer). C7 - immunologic gene sets defined directly from microarray gene expression data from immunologic studies. H - hallmark gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes. & A. BTM - Blood Transcriptional Modules (Chaussabel *et al.*, 2008). B. TISSUES - cell-specific expression from Benita *et al.* (Benita *et al.*, 2010).

Section S5: Classification comparison between DIABLO, Concatenation and Ensemble-based sPLSDA and Elastic net classifiers.

Each integrative classifier was tuned to determine the optimal multi-omics biomarker panel:

- DIABLO models: The tune function in the mixOmics R-library (v6.3.0) was used with a grid of keepX (variables to select on each components) = [2, 5, 10, 15, 20] over 3 components (ncomp=3) either with the null design or full design. A 5x5-fold cross-validation was applied to determine the error rate for various grid value combinations.
    a) When the null design was used (DIABLO_null), the model with the lowest error rate (21%) consisted of 60 mRNA, 42 miRNA and 22 CpGs over 3 components.
    b) When the full design was used (DIABLO_full), the model with the lowest error rate (22%) consisted of 55 mRNA, 17 miRNA and 17 CpGs over 3 components.

  Applying DIABLO_null and DIABLO_full to the test data resulted in an error rate of 19% and 21% respectively.

- Concatenation_sPLSDA: All multi-omics data (mRNA, miRNA and CpGs) were concatenated into one matrix. The tune function in the mixOmics R-library (v6.3.0) was used with a grid of keepX (variables to select on each components) = [2, 5, 10, 15, 20] over 3 components (ncomp=3). A 5x5-fold cross-validation was applied to determine the error rate for various grid value combinations. The model with the lowest error rate (15%) consisting of 60 mRNA but no miRNA or CpGs. Applying Concatenation_sPLSDA to the test data resulted in an error rate of 18%.

- Concatenation_enet: All multi-omics data (mRNA, miRNA and CpGs) were concatenated into one matrix. A grid of lambda values (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1) was used to determine the optimal shrinkage value by applying a 5x5-fold cross-validation using the glmnet R package (v2.0-13). An alpha value of 1 (LASSO penalty) was used to determine a model with the least number of variables. The model with the lowest error rate (14%) consisting of 38 mRNA, 2 miRNA and 118 CpGs. Applying Concatenation_enet to the test data resulted in an error rate of 20%.

- Ensemble_sPLSDA: The tune function in the mixOmics R-library (v6.3.0) was used with a grid of keepX (variables to select on each components) = [2, 5, 10, 15, 20] over 3 components (ncomp=3) and applied to each omics dataset (mRNA, miRNA and CpGs) separately. A 5x5-fold cross-valdiation was used to determine the error rate for each grid value combination for each dataset separately. The model with the lowest error rates for the mRNA, miRNA and CpGs biomarker panels consisted of 60 mRNA, 55 miRNA and 40 CpGs. The cross-validation predictions for these models was combined using an average vote scheme and the resulting error rate for the training data was computed (25%). Appling each model separately to its corresponding data-type and averaging the predictions, resulting in an test error rate of 28%.

- Ensemble_enet: A grid of lambda values (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1) was used to determine the optimal shrinkage value by applying a 5x5-fold cross-validation using the glmnet R package (v2.0-13). The model with the lowest error rates for the mRNA, miRNA and CpGs biomarker panels consisted of 96 mRNA, 45 miRNA and 127 CpGs. The

cross-validation predictions for these models was combined using an average vote scheme and the resulting error rate for the training data was computed (11%). Appling each model separately to its corresponding data-type and averaging the predictions, resulting in an test error rate of 23%.

Section S6: Modular analysis

Eigengene summarization is a common approach to decompose a *n* x *p* dataset (where *n* is the number of samples and *p* is the number of variables in a module), to a component (linear combination of all *p* variables) that represents the summarized expression of genes in the module (Langfelder and Horvath, 2008). For the asthma study, 15,683 genes were reduced to 229 KEGG pathways and 292 metabolites were reduced to 60 metabolic pathways using eigengene summarization.

Section S7: Multilevel transformation

For multivariate analyses, A multilevel approach separates the within subject variation matrix (X_w) and the between subject variation (X_b) for a given dataset (X) (Westerhuis *et al.*, 2010; Liquet *et al.*, 2012), ie. $X = X_w + X_b$. In the case of a two-repeated measured problem (e.g. pre vs post challenge), the within subject variation matrix is similar to calculating the net difference for each individual between the data obtained for pre and post challenge. For each omics dataset, the within-subject variation matrix (X_w) was extracted and used to construct the multilevel DIABLO (mDIABLO) models. In the asthma study, the multilevel approach (called variance decomposition step) was applied to the cell-type, gene and metabolite module datasets.

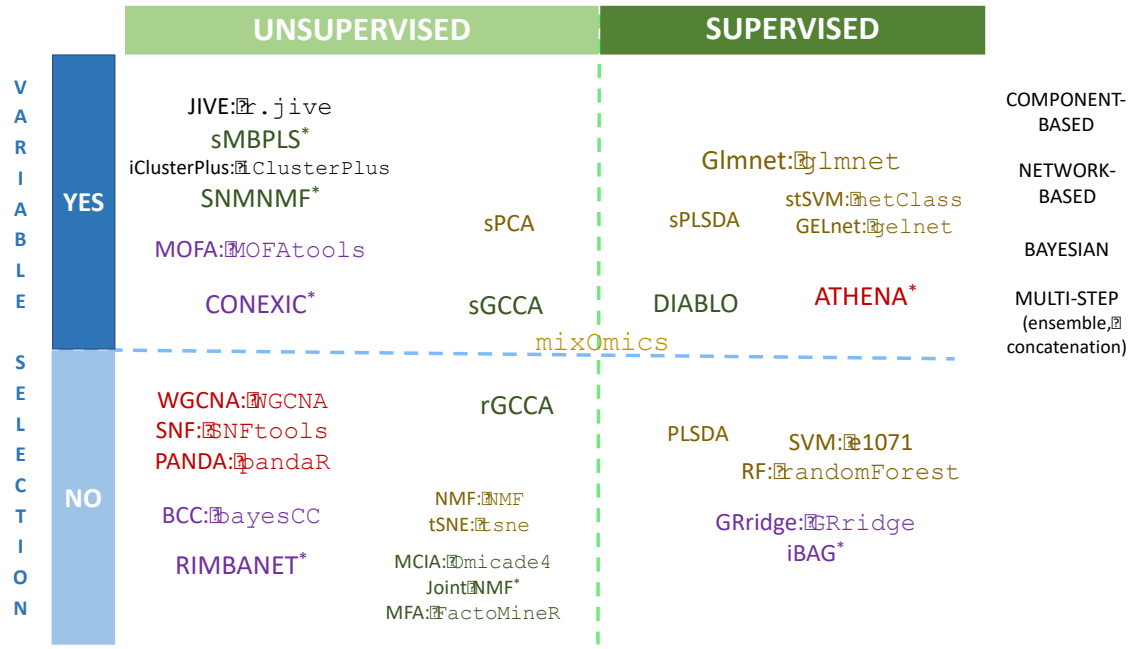

Figure S1. Overview of approaches used for the integration of multiple high dimensional omics datasets using either unsupervised or supervised analyses.

Most integrative methods were developed for unsupervised analyses. Variable selection is an important feature of the methods to improve interpretation of these complex models. Various types of integrative methods are listed, ranging from Component-based that reduce the dimensionality of high-throughput omics datasets, Bayesian methods, Network-based and multi-step approaches which include concatenation and ensemble approaches (Huang *et al.*, 2017). Concatenation-based approach combine multiple matrices and apply standard single omics analysis without taking into account the type of omics variable in the model. Ensemble-based approaches involve the development of independent models for each omics dataset, after which the outputs are combined using various voting schemes (e.g. majority vote, average vote). Methods name in courier font indicate the name of the R package. *Methods are coded in other languages are indicated below.

Abbreviations: JIVE: Joint and Individual Variation Explained (Lock *et al.*, 2013), *sMBPLS: sparse Multiblock Partial Least Squares (Matlab)(Zhang *et al.*, 2012), SNMNMF: Sparse Network-regularized Multiple Non-negative Matrix Factorization (Matlab)(Zhang *et al.*, 2011), MOFA: Multi-Omics Factor Analysis(Argelaguet *et al.*, 2017), *CONEXIC: Copy Number and Expression In Cancer (Java)(An Integrated Approach to Uncover Drivers of Cancer: Cell), WGCNA: Weighted Gene Co-expression Network Analysis(Langfelder and Horvath, 2008), SNF: Similarity Network Fusion(Wang *et al.*, 2014), PANDA: Passing Attributes between Networks for Data Assimilation(Glass *et al.*, 2013), BCC: Bayesian Consensus Clustering(Lock and Dunson, 2013), *RIMBANET: Reconstructing Integrative Molecular Bayesian Networks (Perl)(Zhu *et al.*, 2012); sPCA : sparse Principal Component Analysis(Shen and Huang, 2007); sGCCA: sparse generalized canonical correlation analysis (Tenenhaus *et al.*, 2014); rGCCA: regularized generalized canonical correlation analysis(González *et al.*, 2009); NMF: Non-Negative Factorization (Matlab); MFA: Multiple Co-inertia Analysis (MCIA); Multiple Factor Analysis(Abdi *et al.*, 2013); glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models(Zou and Hastie, 2005); sPLSDA: sparse Partial Least Squares Discriminant Analysis(Lê Cao *et al.*, 2011); stSVM Smoothed t-statistics Support Vector Machine(Cun and Fröhlich, 2013); GELnet: Generalized Elastic Net(Sokolov *et al.*, 2016); *ATHENA: Analysis Tool for Heritable and Environmental Network Associations (Perl)(Kim *et al.*, 2013); SVM: Support Vector Machine; RF: Random Forest(Breiman, 2001); GRridge: Adaptive group-regularized ridge regression(van de Wiel *et al.*, 2016); *iBAG: integrative Bayesian Analysis of Genomics (R and Shiny)(Wang *et al.*, 2013)

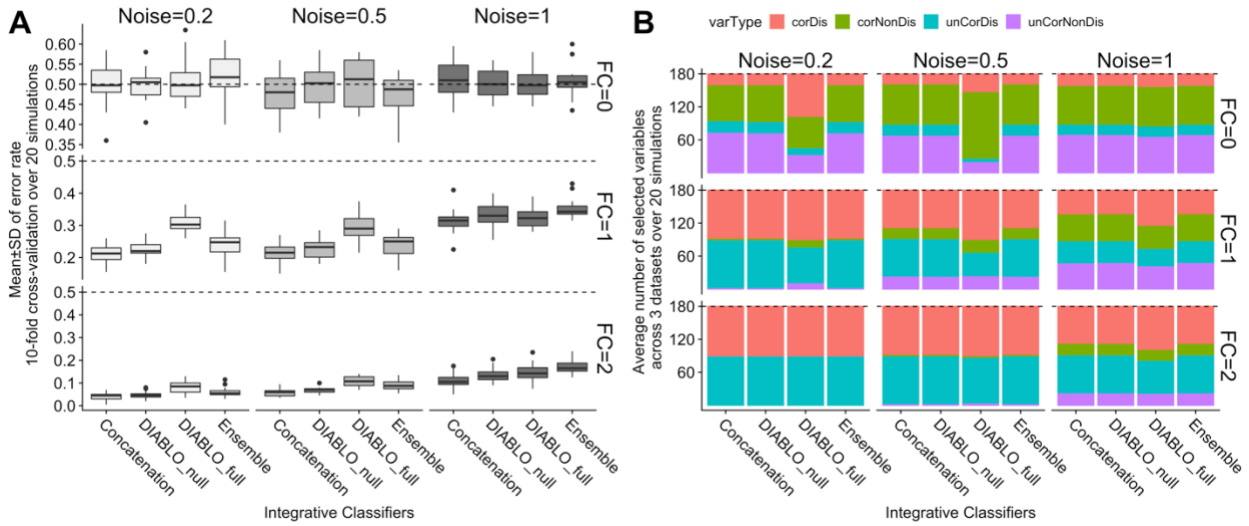

Figure S2. Trade-off between correlation and discrimination in DIABLO models.
A) Classification error rates (10-fold cross-validation averaged over 20 simulations). Dashed line indicates a random performance (error rate = 50%). All methods perform similarly when the fold-change (FC) was zero (first row). All methods performed similarly when the FC=2, that is, the fold-change was greater than the noise and covariance levels. When FC=1, DIABLO_Full had a higher error rate compared to the other methods for noise levels less than 1. B) At lower fold-change levels, DIABLO_Full selected correlated variables (red and green), however, when the fold-change was greater than the noise and covariance levels (FC=2), all methods selected all predictive variables (red, blue).

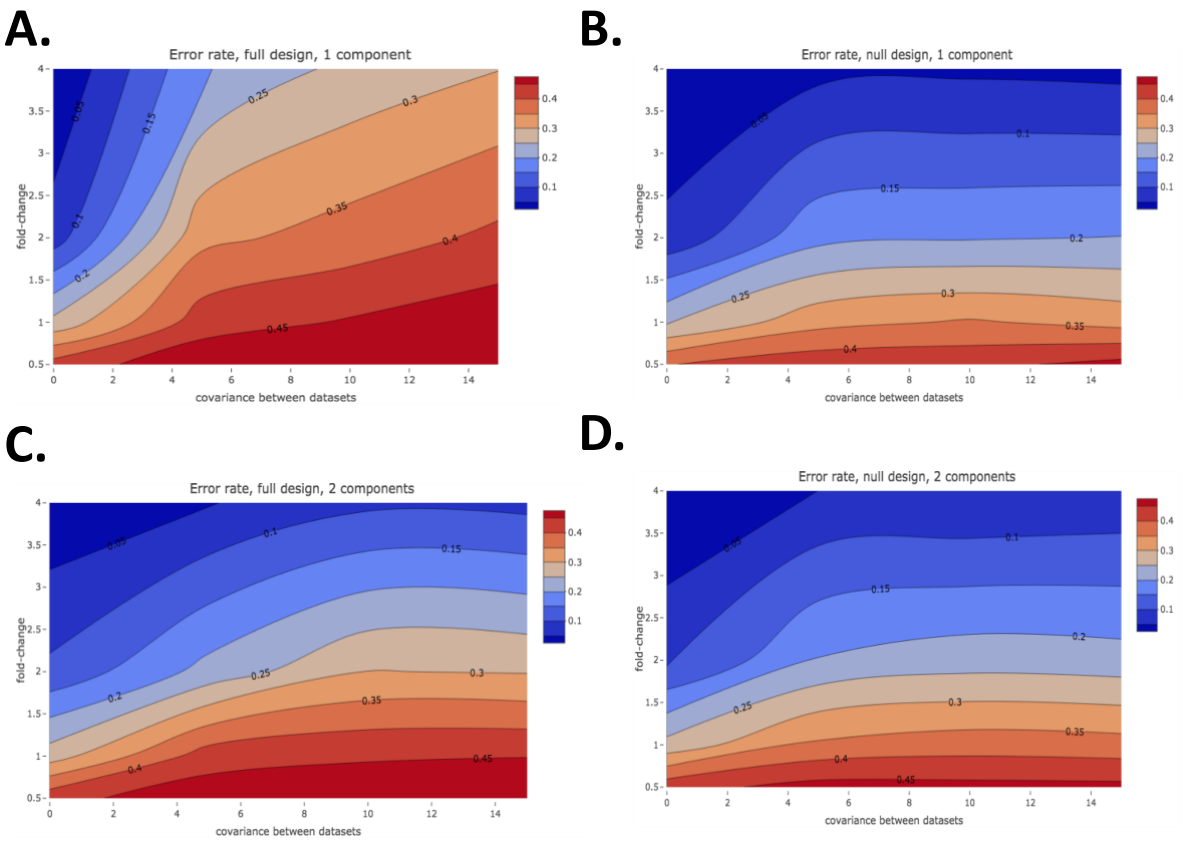

Figure S3. Trade-off between correlation and discrimination: comparison between one or two components.

Contour plots depicting the error rate estimated using 10-fold cross-validation averaged over 20 simulations when the full or null design and retaining either 1 or 2 components. When 1 component is retained, 60 variables were selected per component per dataset whereas when 2 components were selected 30 variables were selected per component per dataset. Therefore all DIABLO models consisted of 180 variables (60 variables per dataset). Increasing the covariance between datasets significantly increased the error rate for a given fold-change (blue to red) for the DIABLO_Full model (A) as compared to the DIABLO_Null model (B). The error rates between the DIABLO_Full and DIABLO_Null models are more comparable when 2 components are retained (C-D).

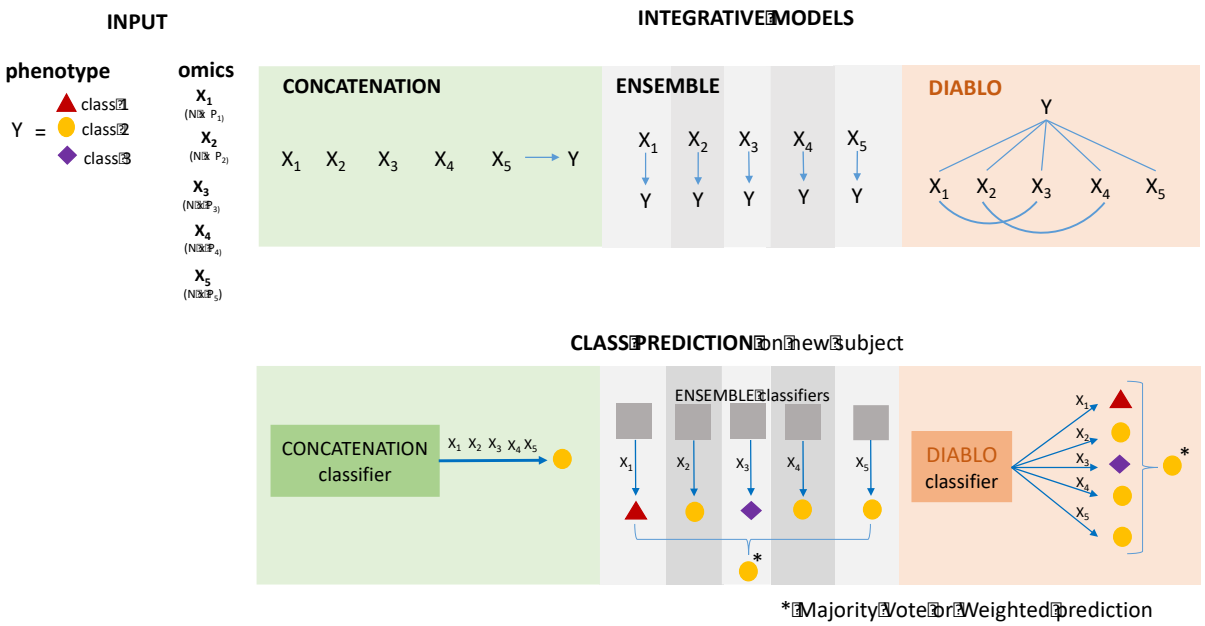

Figure S4. Integrative prediction frameworks including multi-step approaches (concatenation, ensemble) and DIABLO to identify multi-omics molecular signatures.

Concatenation-based integration combines multiple datasets into a single large dataset, with the aim to predict a phenotype of interest. Ensemble-based classification methods construct a predictive model on each individual dataset before combining the model predictions. None of these approaches account or model relationships between datasets and thus limit our understanding of molecular interactions at multiple functional levels. DIABLO simultaneously maximizes the associations between datasets and a phenotype of interest to identify a correlated set of variables of different omics-types that are also discriminatory. The prediction is based on each omics-associated component derived from the model. All methods presented here are data-driven approaches, which do not use any prior knowledge such as from curated biological databases (eg. protein-protein interactions).

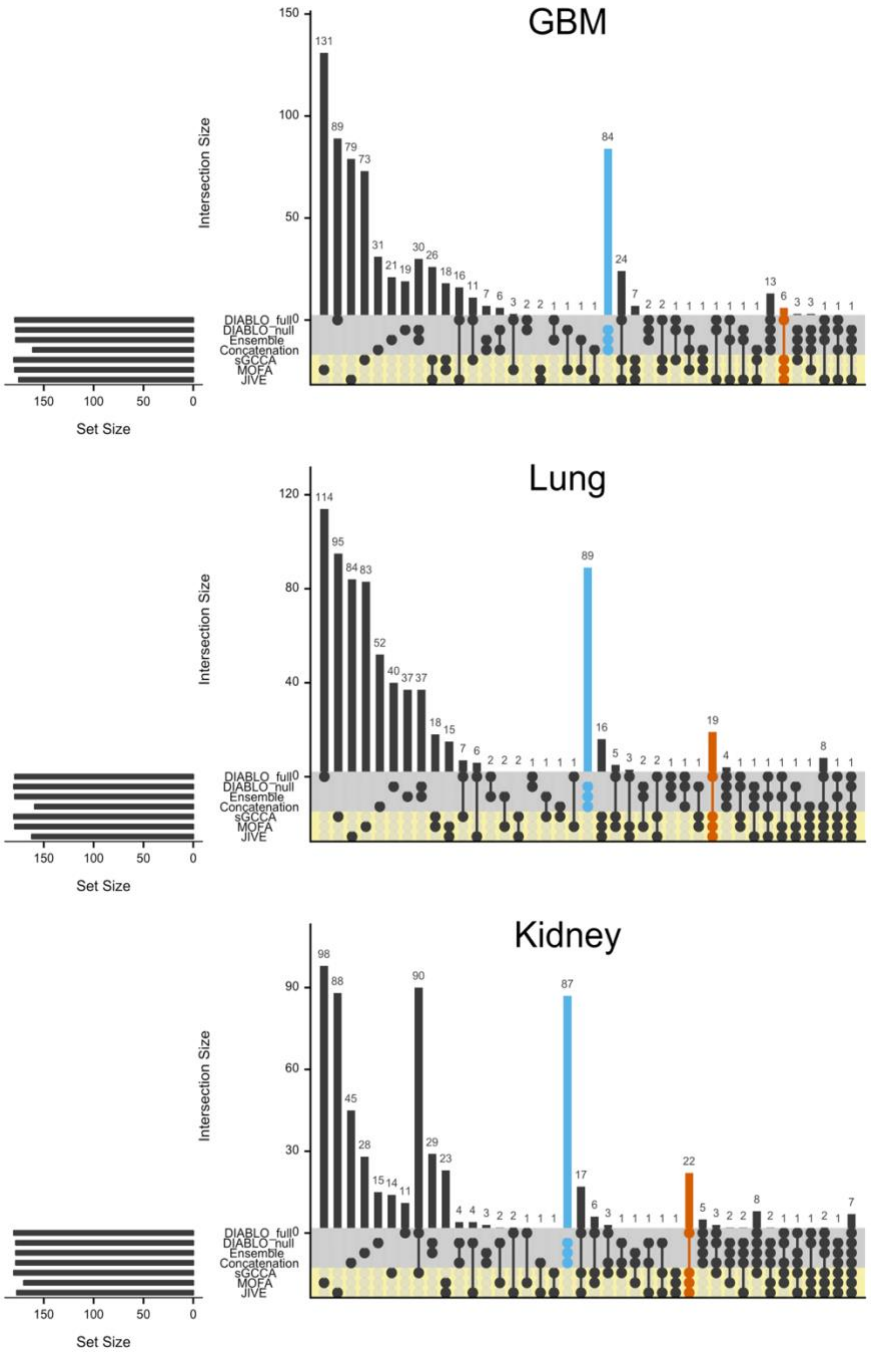

Figure S5. Benchmark analyses: overlap between multi-omics biomarker panels.
Intersection plots of multi-omics biomarker panels identified using both supervised (gray) and unsupervised (yellow) methods for the gbm, kidney and lung cancer datasets. For each method 2 components were retained, and 30 variables were selected for each dataset, resulting in 30 variables x 2 components x 3 datasets = 180 variables per method. Although the first and second components are orthogonal to each other, some variables were selected on both components. The set size depicts the number of unique features and thus leads to the unequal set size depicted above. The

largest overlap is often observed between the supervised methods, with the exception of DIABLO_full (blue bar), which was more similar to unsupervised methods (orange bar).

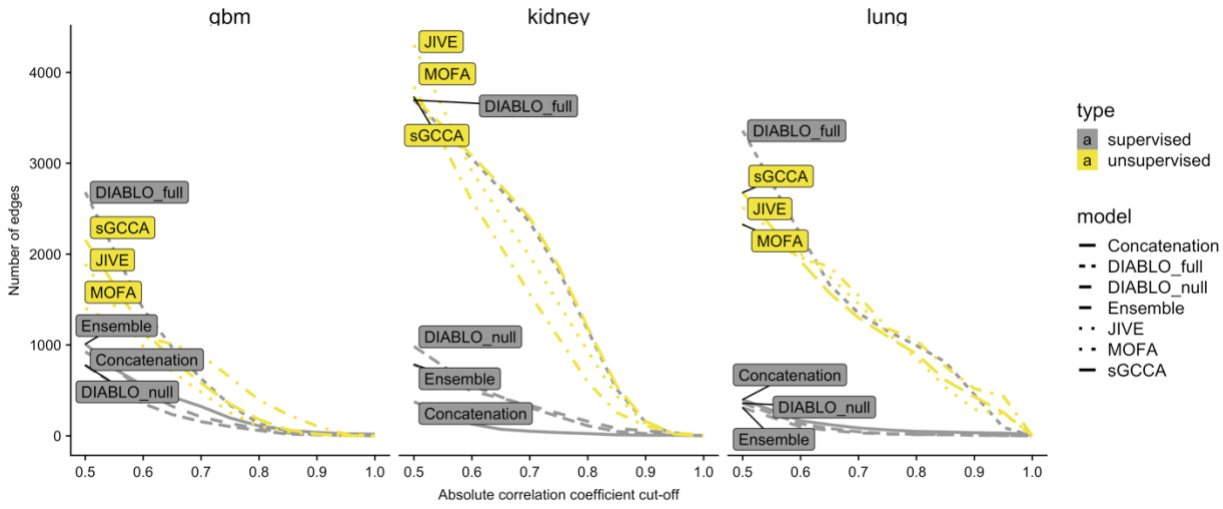

Figure S6. Benchmark analyses: Number of correlated variables at various correlation cut-offs.
A correlation matrix was computed using the variables select in the multi-omics biomarker panel identified for each multi-omics cancer datasets. At various correlation coefficient cut-offs, the number of features that were correlated with other features is depicted for panels identified using both supervised and unsupervised methods. The unsupervised methods lead to a higher number of connections (edges) irrespective of the correlation cut-off, as compared to the supervised methods, with the exception of DIABLO_full.

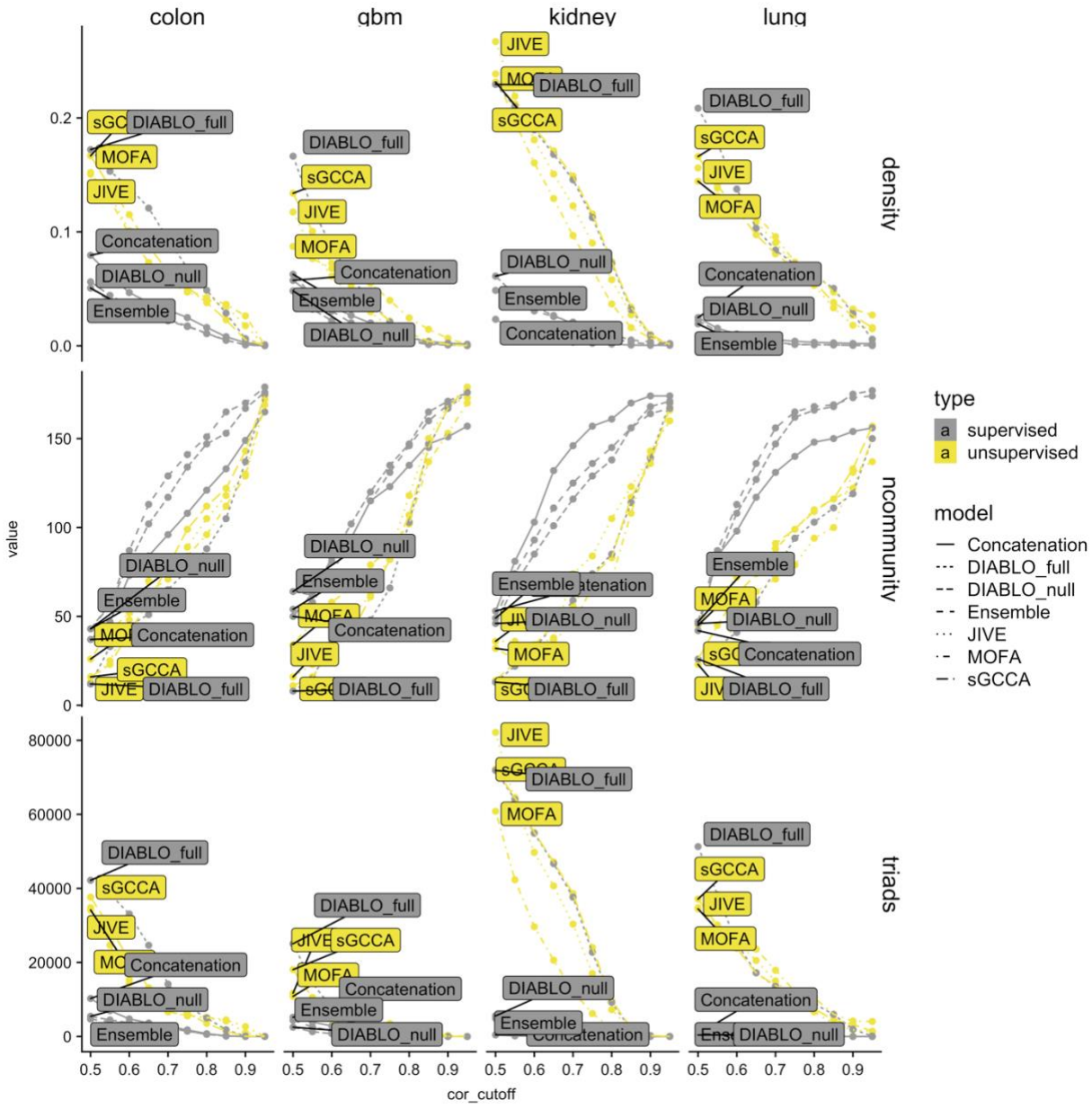

Figure S7. Benchmark analyses: network properties of multi-omics signatures.
We analysed each of the four multi-omics cancer datasets with component-based integrative methods with variable selection. The network attributes, density, number of communities and triads resulting from each molecular signature are represented. The unsupervised methods (yellow) led to multi-omics signatures with a higher graph density, a greater number of triads and a lower number of communities as compared to supervised methods (gray), with the exception of DIABLO_full which simultaneously explained the correlation structure between multiple omic datasets and a phenotypic response variable.

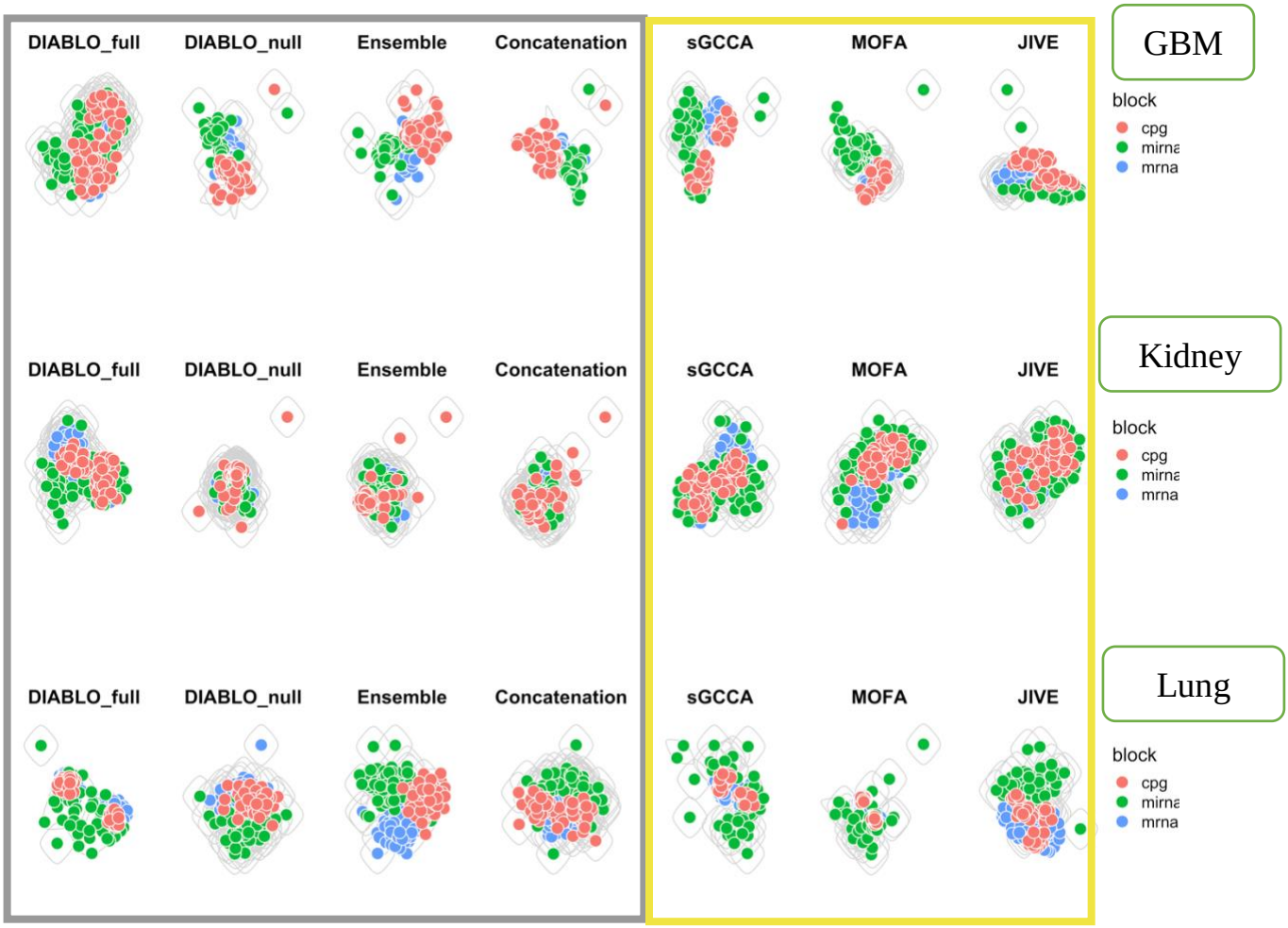

Figure S8. Benchmark analyses: network connectivity of multi-omics signatures.
Networks of the multi-omics biomarker panels identified from each method are represented for a Pearson's correlation cut-off of |0.4|. The edge betweenness as computed to estimate the number of modules (depicted by the gray circles).

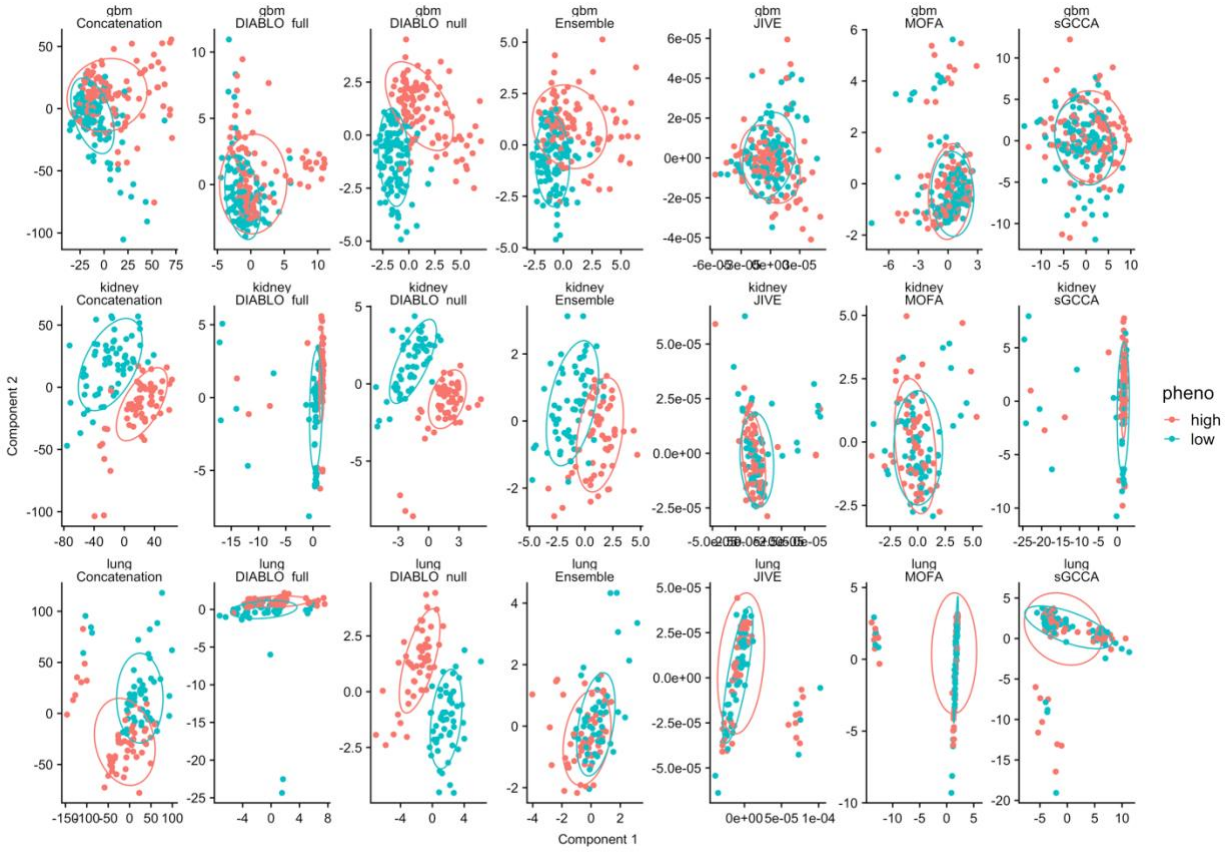

Figure S9. Benchmark analyses: sample plots for each multi-omics panel.

As expected, a strong separation between high and low survival groups can be observed for supervised methods but not for unsupervised methods. The level of discrimination decreases when using DIABLO_full as compared to DIABLO_null.

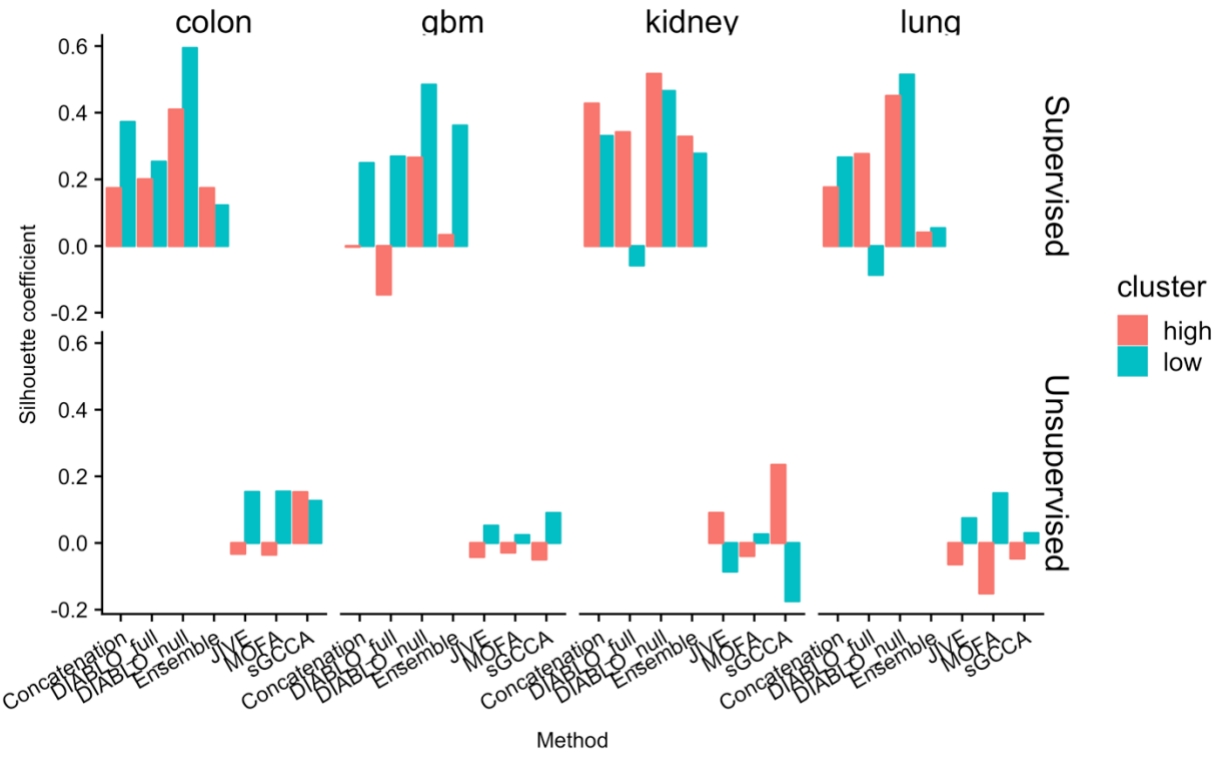

Figure S10. Internal validation of high and low phenotypic groups for all method in the benchmarking experiments.

The silhouette for each data *i*, was computed as the normalized difference between two average distances (a_i and b_i), where a_i is the average distance between *i* and all points within its own cluster and *bi* is the average distance between *i* and all points that are not in its cluster ($s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$). The silhouette ranges from -1 to 1, 1 being a strong indicator of cluster membership and -1 being a weak indicator of cluster membership. As can be observed, the supervised methods show stronger silhouette coefficients as compared to unsupervised methods. This is because the principal components are associated with the phenotype of interest. DIABLO_Null consistently out-performed the methods with a higher average silhouette coefficient with respect to both phenotypic groups (high and low survival). The silhouette coefficients for the other methods were variable, however, whether this translates to a lower predictive performance in independent test data remains to be observed.

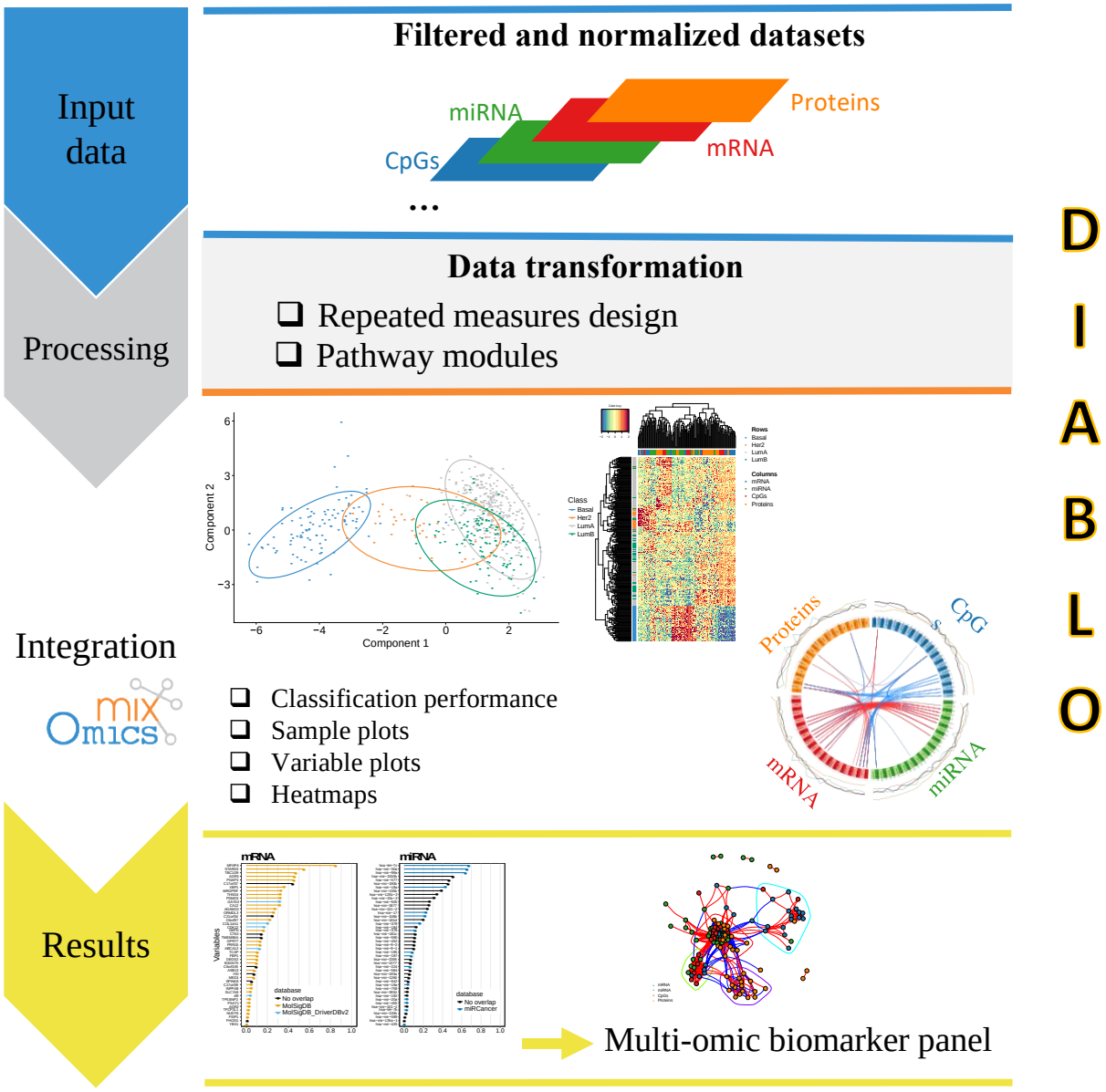

Figure S11. A standard DIABLO workflow.
The first step inputs multiple omics datasets measured on the same individuals, that were previously normalized and filtered, along with the phenotype information indicating the class membership of each sample (two or more groups). Optional preprocessing steps include multilevel transformation for repeated measures study designs and pathway module summary transformations. DIABLO is a multivariate dimension reduction method that seeks for latent components – linear combinations of variables from each omics dataset, that are maximally correlated as specified by a design matrix (see Methods section). The identification of a multi-omics panel is obtained with l_1 penalties in the model that shrink the variable coefficients defining the components to zero. Numerous visualizations are proposed to provide insights into the multi-omics panel and guide the interpretation of the selected omics variables, including sample and variable plots. Downstream analysis includes gene set enrichment analysis.

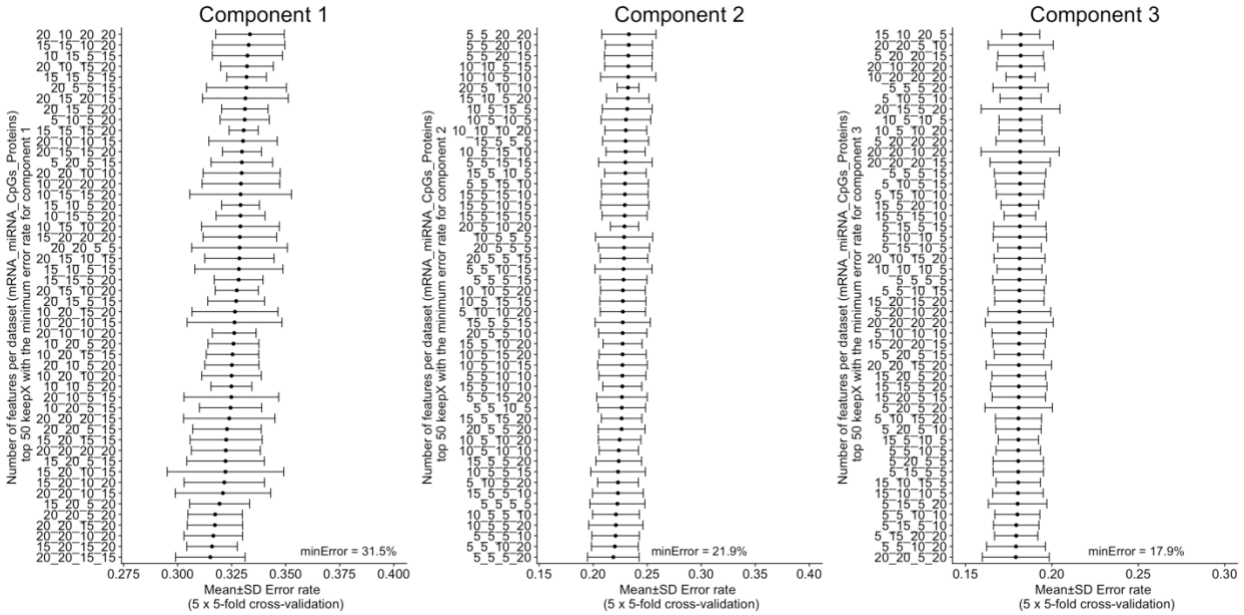

Figure S12. Breast cancer multi omics study: optimal multi-omics biomarker panel for PAM50 subtypes.

A grid was used to identify the optimal combination of variables select from each omics datasets. The following grid values was used for each omics dataset: mRNA = [5, 10, 15, 20], miRNA = [5, 10, 15, 20], CpGs = [5, 10, 15, 20], Proteins = [5, 10, 15, 20], across 3 components. The centroids distance measure was used to compute the error rate (Rohart *et al.*, 2017). The optimal multi-omics panel consisted of 20 mRNAs, 20 miRNAs, 15 CpGs and 15 proteins on component 1, 5 mRNAs, 5 miRNAs, 5 CpGs and 20 proteins on component 2, and 20 mRNAs, 20 miRNAs, 5 CpGs and 20 proteins on component 3.

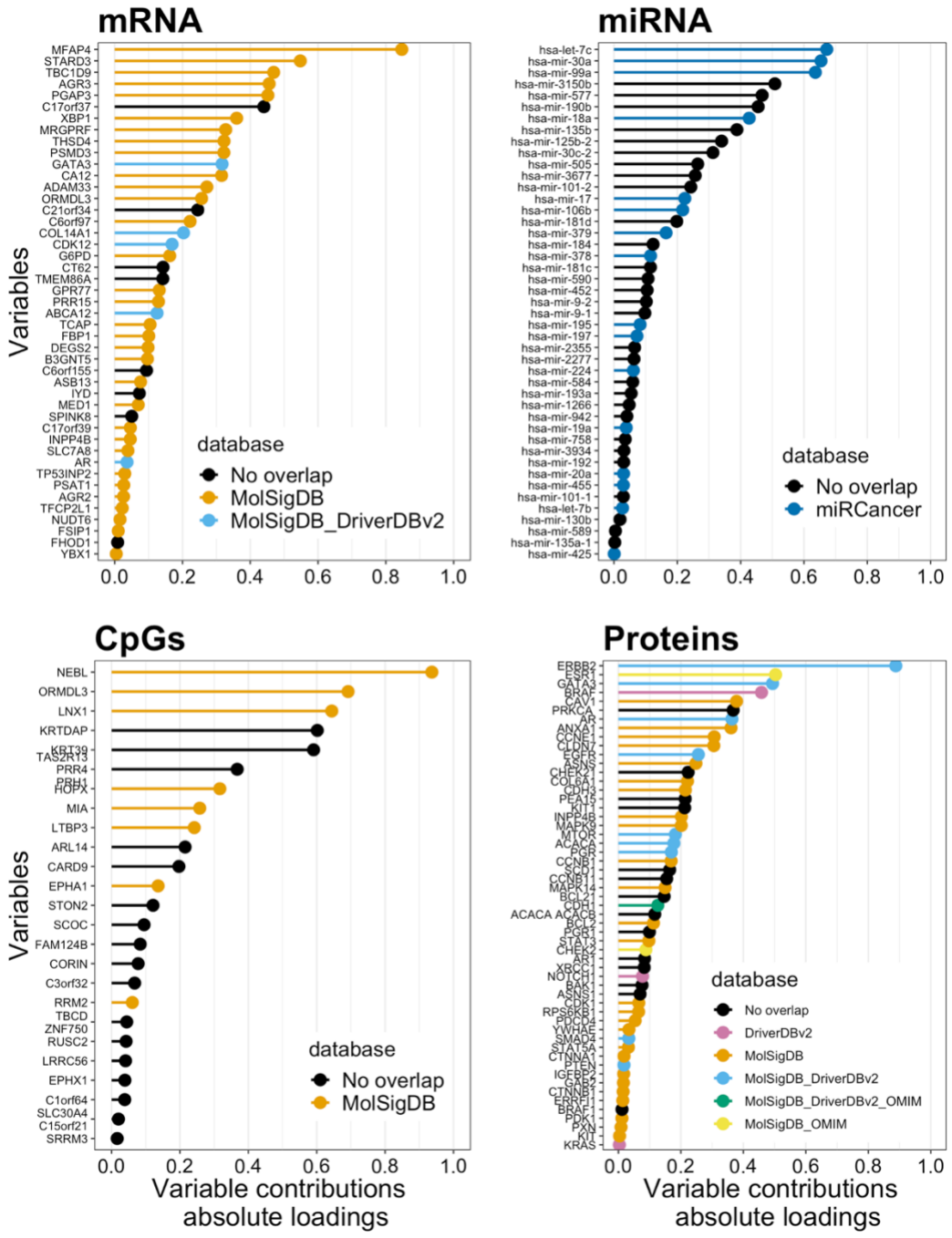

Figure S13. Variable importance plots for the breast cancer multi-omics biomarker panel.
The variable importance based on the absolute value of the weights on the loading vectors were plotted for each omic-type as part of the multi-omics biomarker panel predictive of PAM50 breast cancer subtypes identified using DIABLO_full. Each variable is color-code based on its existence in databases that associate variables with breast cancer. Variables in black have no known associations in curated biological datasets with respect to breast cancer.

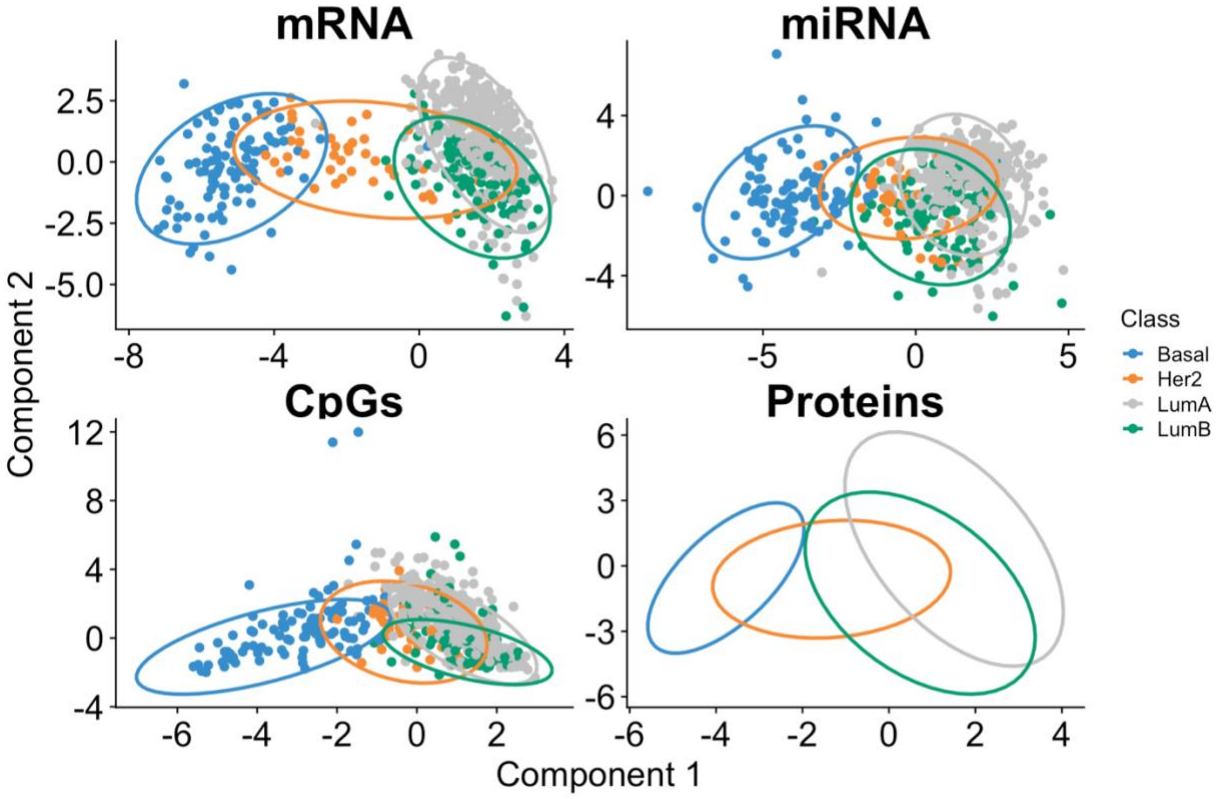

Figure S14. Omic-specific component plots.
Component plots for each omic dataset depicting the clustering of subjects with respect to the PAM50 subtypes. The 95% confidence ellipses are based on the training model and superimposed with test data.

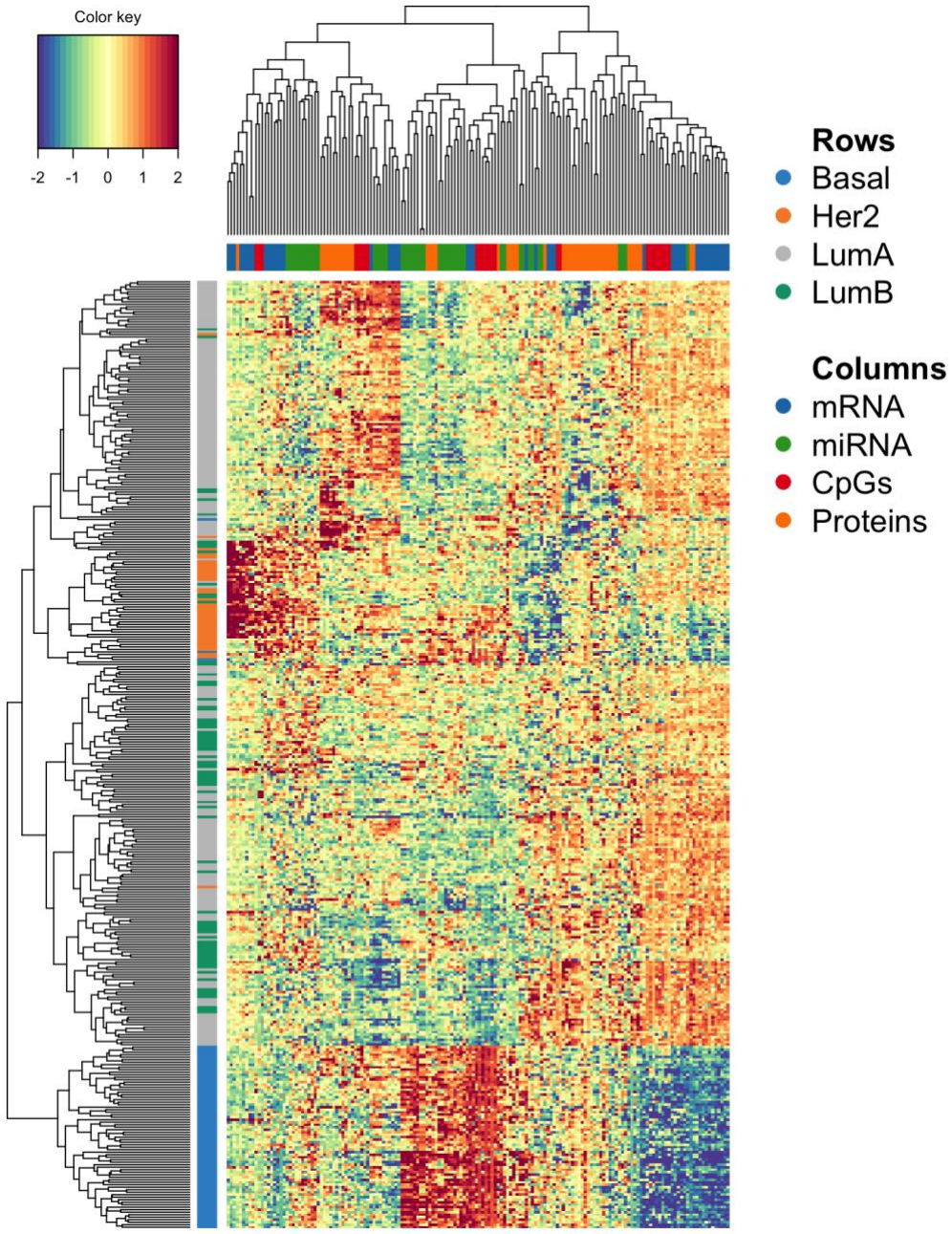

Figure S15. Heatmap of scaled expression of the variables identified in the multi-omics biomarker panels.

The expression values of all variables that were part of the multi-omics biomarker panel identified to be predictive of PAM50 breast cancer subtypes were scaled and underwent hierarchical clustering. As can be observed samples with the Her2 and Basal subtypes cluser strongly whereas LumA and LumB are much harder to separate from each other.

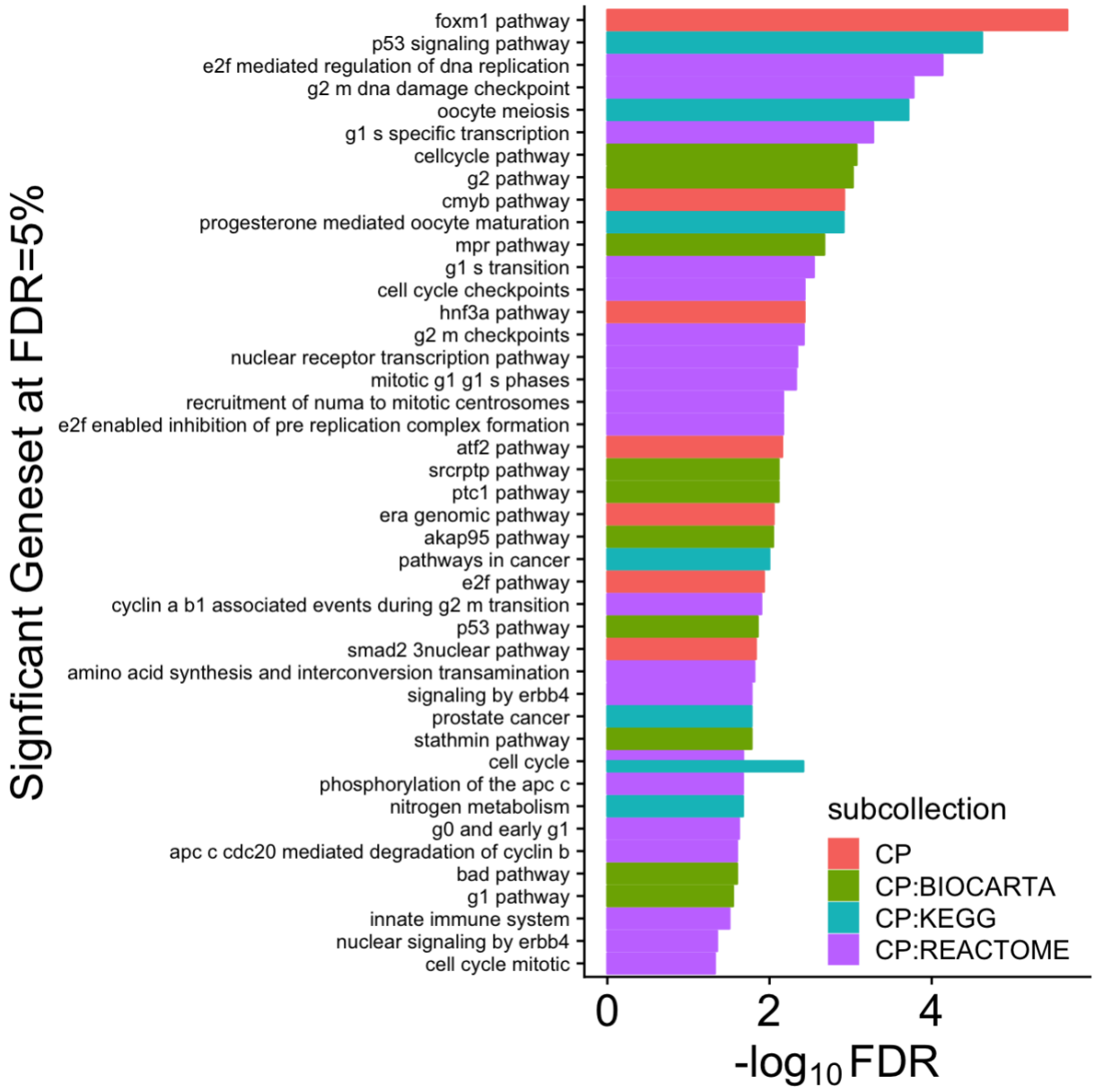

Figure S16. Significant pathways enriched in the largest community identified using the features of multi-omics biomarker panel for PAM50 subtypes.

The largest cluster (in Figure 3B) consisted of 72 variables; 20 mRNAs, 21 miRNAs, 15 CpGs and 16 proteins (red bubble) and was further investigated using gene set enrichment analysis. The barchart depicts the enriched genesets at an FDR cut-off of 5%.

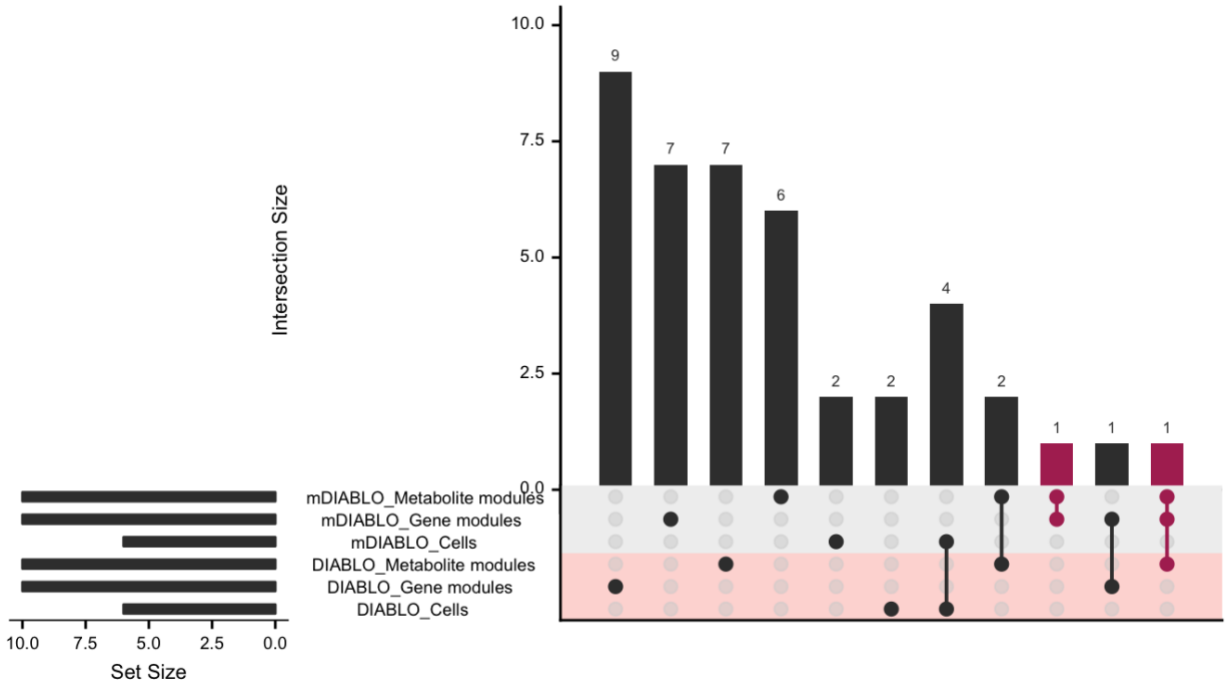

Figure S17. Overlap between biomarker panels identified using DIABLO and multilevel DIABLO. The intersection (overlap) between variables selected by applying mDIABLO and the standard DIABLO model. Only mDIABLO identified variables that spanned different biological domains (red).

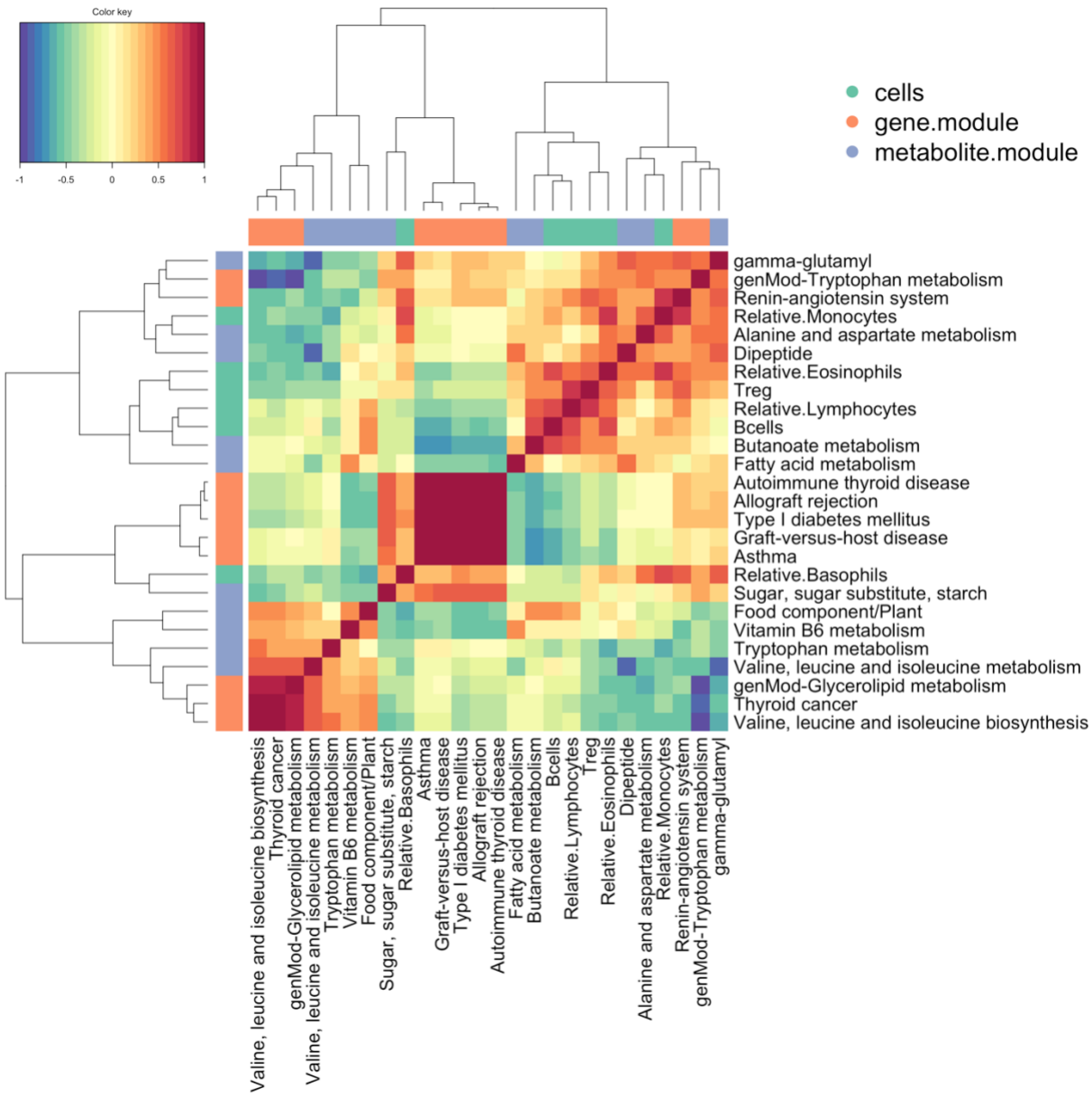

Figure S18. Heatmap depicting the correlation matrix of the variables identified using multilevel DIABLO (mDIABLO).

The correlation matrix computed based on the features selected by mDIABLO depicts strong groups of highly correlated features.

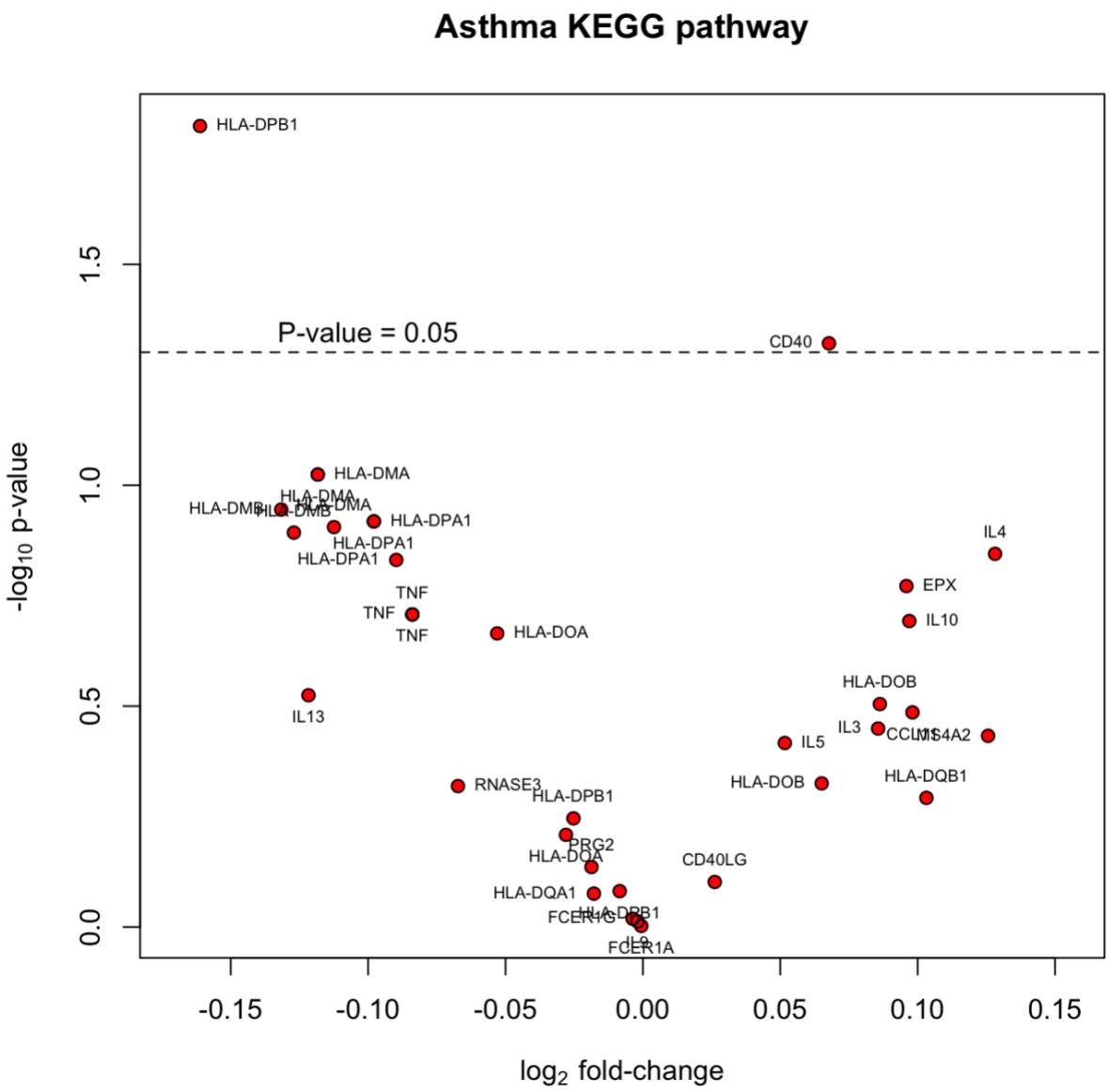

Figure S19. Asthma multi-omics study: volcano plot of genes in the Asthma KEGG pathway.
The volcano plot depicts the significance of each gene in the asthma pathways against its respective fold-change (change in expression from pre to-post challenge). The significance is based on a paired *t*-test. The volcano plot shows that with the exception of HLA-DPB1 and CD40 no other genes within the Asthma pathway were significant at the nominal p-value cut-off of 0.05. However, this pathway was selected by DIABLO as a strong predictor of allergen challenge. This modular-based analysis depicts the power of combining genes with small effect sizes which together contribute to a pathway that significantly changes in response to allergen inhalation challenge.

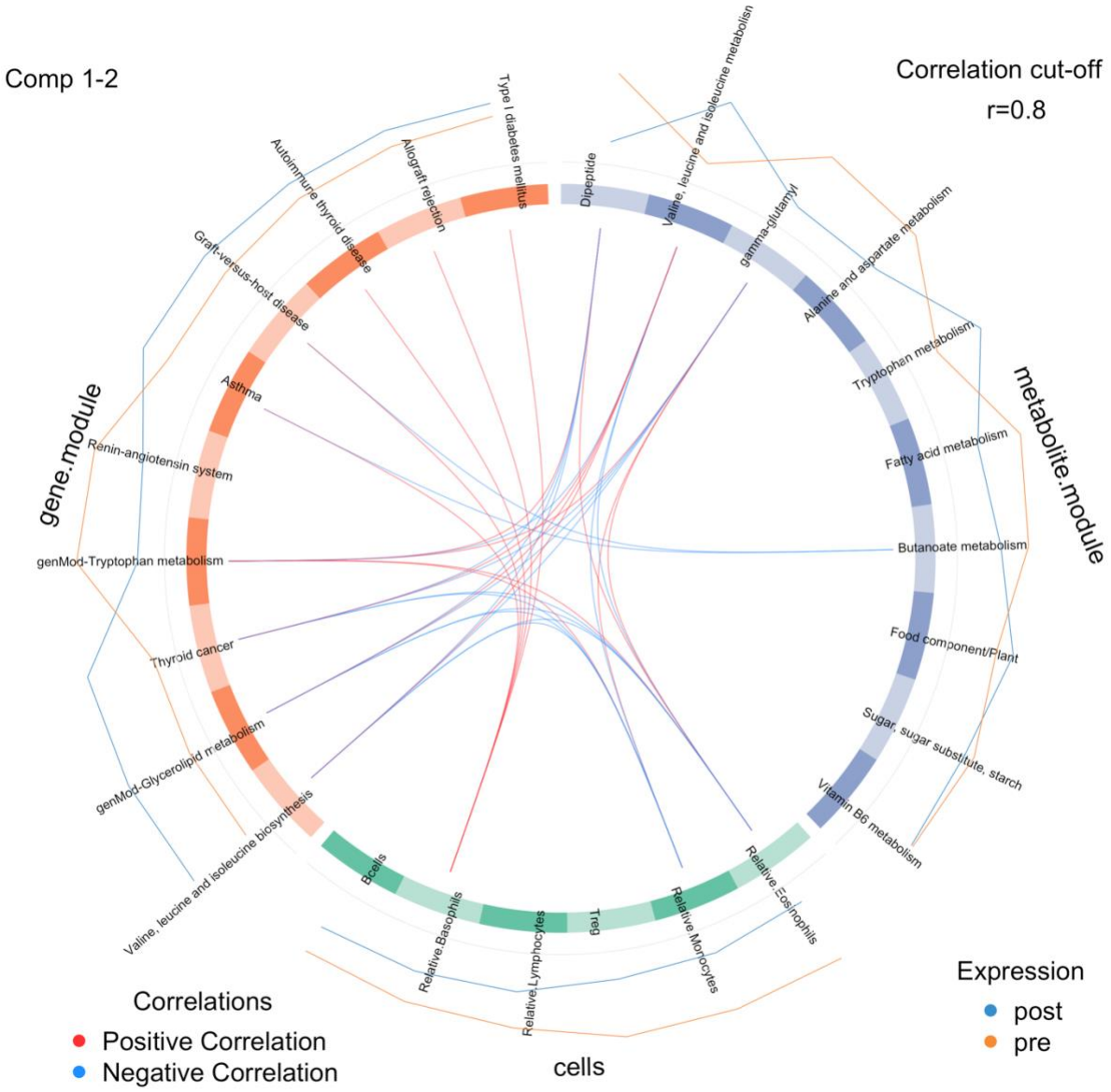

Figure S20. Circos plot depicting the strongest correlation biomarkers in the multi-omics biomarker panel.

The variables selected by applying mDIABLO to cellular frequencies, gene and metabolite module datasets are depicted using a circos plot. The variables are indicated in the ideogram and connected with either red or blue to other variables if the correlation is either positive or negative. Only correlation above a certain threshold are depicted (r=0.8). The lines around the ideogram are drawn by connecting the average expression value of a given variable for a certain phenotypic group.

Table S1. Number of significant gene sets for each integrative method and benchmarking cancer dataset.

Best performing method is indicated in the shaded cell. Each row represents a gene set collection (see **Suppl. Section S4** for details, FDR = 5%).

		Unsupervised, integrative			**Supervised, non-integrative**			**Supervised, integrative**
disease	**collection**	**JIVE**	**MOFA**	**sGCCA**	**Concatenation**	**Ensemble**	**DIABLO_null**	**DIABLO_full**
Colon	**BTM**	0	4	0	0	0	0	23
	C1	0	0	0	0	0	0	0
	C2	15	14	5	12	3	21	113
	C3	8	5	14	11	2	6	0
	C4	0	1	0	1	2	1	46
	C5	19	36	147	7	0	0	216
	C6	0	0	0	0	0	0	0
	C7	1	87	11	61	10	62	218
	H	0	0	0	0	0	2	7
	TISSUES	2	12	0	0	0	0	16
	TOTAL	45	159	177	92	17	92	639
Gbm	**BTM**	0	0	19	10	9	10	30
	C1	0	0	0	0	0	0	0
	C2	275	337	193	258	358	312	426
	C3	94	64	37	14	15	15	34
	C4	49	43	68	47	50	62	125
	C5	825	708	706	526	669	776	693
	C6	22	25	18	30	24	24	21
	C7	460	82	526	432	173	147	869
	H	12	8	8	19	23	20	19
	TISSUES	18	29	21	10	12	14	44
	TOTAL	1755	1296	1596	1346	1333	1380	2261
Kidney	**BTM**	1	0	0	0	0	0	0
	C1	0	0	1	0	0	0	1
	C2	42	33	7	10	5	15	4
	C3	8	80	1	4	35	23	1
	C4	17	6	0	7	1	3	0
	C5	157	110	1	55	27	46	0
	C6	0	0	0	0	0	0	0
	C7	0	74	15	93	13	10	18
	H	6	3	0	1	0	1	0
	TISSUES	2	0	0	0	0	0	0

	TOTAL	233	306	25	170	81	98	24
Lung	**BTM**	0	0	0	0	0	2	0
	C1	0	0	0	1	0	0	1
	C2	4	17	2	0	0	1	33
	C3	48	20	57	50	26	21	19
	C4	17	0	47	0	0	18	13
	C5	35	127	42	0	25	22	193
	C6	1	0	1	3	2	5	7
	C7	18	13	78	0	7	72	100
	H	0	2	0	0	1	0	0
	TISSUES	0	0	0	0	0	9	20
	TOTAL	123	179	227	54	61	150	386

Table S2. Classification error rates (average error, sd) of DIABLO, Concatenation-based and Ensemble-based sPLSDA and Elastic Net (enet) classifiers on the Breast Cancer study (see Suppl. Section S5 for details).

Dataset	***p***	**Train**	**Test**
Diablo_null	mRNA: 60 miRNA: 42 CpGs: 22	0.21 (0.0091)	0.19
Diablo_full	mRNA: 55 miRNA: 17 CpGs: 17	0.22 (0.0057)	0.21
Concatenation_sPLSDA	mRNA: 60 miRNA: 0 CpGs: 0	0.15 (0.013)	0.18
Concatenation_enet	mRNA: 38 miRNA: 2 CpGs: 118	0.14 (0.0072)	0.20
Ensemble_sPLSDA	mRNA: 60 miRNA: 55 CpGs: 40	0.25 (0.014)	0.28
Ensemble_enet	mRNA: 96 miRNA: 45 CpGs: 127	0.11 (0.0016)	0.23

References

Abdi,H. *et al.* (2013) Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdiscip. Rev. Comput. Stat.*, **5**, 149–179.

An Integrated Approach to Uncover Drivers of Cancer: Cell.

Argelaguet,R. *et al.* (2017) Multi-Omics factor analysis disentangles heterogeneity in blood cancer. *bioRxiv*, 217554.

Argelaguet,R. *et al.* (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, **14**, e8124.

Benita,Y. *et al.* (2010) Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. *Blood*, **115**, 5376–5384.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, 289–300.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chaussabel,D. *et al.* (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*, **29**, 150–164.

Cun,Y. and Fröhlich,H. (2013) Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PLoS ONE*, **8**, e73074.

Glass,K. *et al.* (2013) Passing messages between biological networks to refine predicted interactions. *PLoS ONE*, **8**, e64832.

González,I. *et al.* (2009) Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *J. Biol. Syst.*, **17**, 173–199.

Huang,S. *et al.* (2017) More is better: recent progress in multi-omics data integration methods. *Front. Genet.*, **8**.

Kim,D. *et al.* (2013) ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Min.*, **6**, 23.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Law,C.W. *et al.* (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, **15**, R29.

Lê Cao,K.-A. *et al.* (2011) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, **12**, 253.

Liberzon,A. *et al.* (2015) The molecular signatures database hallmark gene set collection. *Cell Syst.*, **1**, 417–425.

Liquet,B. *et al.* (2012) A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinformatics*, **13**, 325.

Lock,E.F. *et al.* (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.*, **7**, 523–542.

Lock,E.F. and Dunson,D.B. (2013) Bayesian consensus clustering. *Bioinformatics*, **29**, 2610–2616.

Rohart,F. *et al.* (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.*, **13**, e1005752.

Shen,H. and Huang,J. (2007) Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation. *J. Multivar. Anal.*, **99**, 1015–1034.

Singh,A. *et al.* (2013) Gene-metabolite expression in blood can discriminate allergen-induced isolated early from dual asthmatic responses. *PLoS ONE*, **8**, e67907.

Singh,A. *et al.* (2012) Plasma proteomics can discriminate isolated early from dual responses in asthmatic individuals undergoing an allergen inhalation challenge. *PROTEOMICS - Clin. Appl.*, **6**, 476–485.

Singh,A. *et al.* (2014) Th17/Treg ratio derived using DNA methylation analysis is associated with the late phase asthmatic response. *Allergy Asthma Clin. Immunol.*, **10**, 32.

Sokolov,A. *et al.* (2016) Pathway-based genomics prediction using generalized elastic net. *PLoS Comput Biol*, **12**, e1004790.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, **102**, 15545–15550.

Tenenhaus,A. *et al.* (2014) Variable selection for generalized canonical correlation analysis. *Biostatistics*, **15**, 569–583.

Wang,B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Wang,W. *et al.* (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, **29**, 149–159.

Westerhuis,J.A. *et al.* (2010) Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics*, **6**, 119–128.

van de Wiel,M.A. *et al.* (2016) Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat. Med.*, **35**, 368–381.

Zhang,S. *et al.* (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, **27**, i401–i409.

Zhang,S. *et al.* (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.

Zhu,J. *et al.* (2012) Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol.*, **10**, e1001301.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 301–320.